

# SHAPE FEATURES IMPROVE GENERAL MODEL ROBUSTNESS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Recent studies show that convolutional neural networks (CNNs) are vulnerable under various settings, including *adversarial examples*, *distribution shifting* and *backdoor attacks*. Motivated by the findings that human visual system pays more attention to global structure (e.g., shape) for recognition while CNNs are biased towards local texture features in images, we propose a unified framework `EdgeGANRob` based on robust edge features to improve the robustness of CNNs on multiple tasks, which first explicitly extracts shape/structure features from a given image and then reconstructs a new image by refilling the texture information with a trained generative adversarial network (GAN). In addition, to reduce the sensitivity of edge detection algorithm to adversarial perturbation, we propose a robust edge detection approach *Robust Canny* based on the vanilla Canny algorithm. To gain more insights, we also compare `EdgeGANRob` with its simplified backbone procedure `EdgeNetRob`, which performs learning tasks directly on the extracted robust edge features. We find that `EdgeNetRob` can help boost model robustness significantly but at the cost of the clean model accuracy. `EdgeGANRob`, on the other hand, is able to improve clean model accuracy compared with `EdgeNetRob` and without losing the robustness benefits introduced by `EdgeNetRob`. Extensive experiments show that `EdgeGANRob` is resilient in different learning tasks under diverse settings.

## 1 INTRODUCTION

Convolutional neural networks (CNNs) have been studied extensively (Goodfellow et al., 2016), and have achieved state-of-the-art performance in many learning tasks (He et al., 2016; Zhu et al., 2017). However, recent works have shown that CNNs are vulnerable to *adversarial examples* (Carlini and Wagner, 2017; Goodfellow et al., 2014b; Szegedy et al., 2013), where imperceptible perturbation can be added to the test data to tamper the predictions. Different from *adversarial examples* where test data is manipulated, an orthogonal setting: *data poisoning* or *backdoor attacks* where training data is manipulated to reduce model’s generalization accuracy and achieve targeted poisoning attack (Li et al., 2016; Chen et al., 2017b). In addition, recent studies show that CNNs tend to learn surface statistical regularities instead of high level abstraction, leading it fails to generalize to the superficial pattern transformation (radial kernel, random kernel (Jo and Bengio, 2017a; Wang et al., 2019a;b)). We refer to this problem as model’s robustness under *distribution shifting*. How to improve the general robustness of DNNs under these settings remains unsolved.

To improve the robustness of CNNs, recent studies explore the underlying cause of their vulnerability. For example, Ilyas et al. (2019) attributes the existence of adversarial examples to the existence of non-robust but highly-predictive features. They suggest to train a classifier only on “robust features” which contain the necessary information for recognition and are insensitive to small perturbations. In addition, it is shown that human recognition relies mainly on global object shapes rather than local patterns (e.t. textures), while CNNs are more biased towards the latter (Baker et al., 2018; Geirhos et al., 2019). For instance, Geirhos et al. (2019) creates a texture-shape cue conflict, such as a cat shape with elephant texture, and feeds it to an Imagenet trained CNN and huamn respectively. While Human can still recognize it as a cat, CNN wrongly predicts it as an elephant. Therefore, the bias toward local features potentially contributes to CNN’s vulnerability to adversarial examples, distribution shifting and patterns of backdoor attacks. Particularly, previous researcher also shows

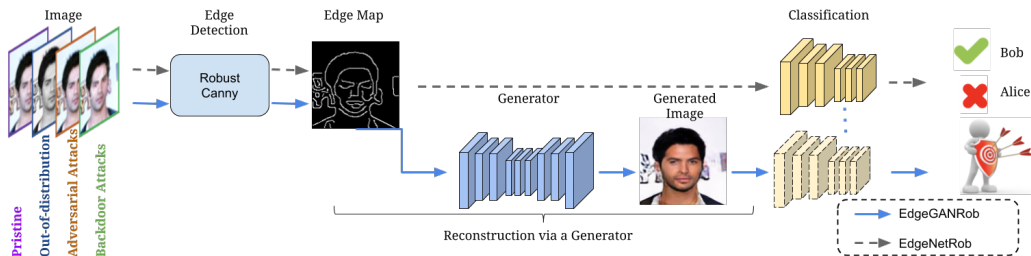


Figure 1: Structure of the proposed pipeline. EdgeNetRob feeds the output of edge detection to the classifier to produce robust predictions, while EdgeGANRob refill the edge image with texture information to reconstruct a new instance for predictions.

that the shape of objects is the most important cue for human object recognition (Landau et al., 1988).

Given the above evidence, a natural question emerges: Can we improve the robustness of CNNs by making it rely more on global shape structure? To answer this question, we need to formalize the notion of global shape structure first. We propose to consider a specific type of shape representation: edges (image points that have sharp change in brightness). Using edges comes with two benefits: 1) it is an effective device for modelling shape; 2) edges are easy to be captured in images, with many sophisticated algorithms (Canny, 1986; Xie and Tu, 2015; Liu et al., 2017) available.

More specifically, this paper explores a new approach EdgeGANRob to improve the robustness of the CNNs to *adversarial attacks*, *distribution shifting* and *backdoor attacks* by leveraging structural information in images. The unified framework is shown in Figure 1. As illustrated, a simplified version of EdgeGANRob is a two-stage procedure named EdgeNetRob, which extracts the structural information by detecting edges and then trains the classifier on the extracted edges. As a consequence, EdgeNetRob forces the CNNs to make prediction solely based on shape information, rather than texture/color, thus eliminating the texture bias (Geirhos et al., 2019). Our results show that EdgeNetRob can improve CNNs’ robustness. However, there are still two challenges: (i) the direct differentiable edge detection algorithms are also vulnerable to attacks, which may lead to low robustness against sophisticated adaptive attackers. To handle this problem, we propose a robust edge detection algorithm, *Robust Canny*. Using *Robust Canny* is able to EdgeNetRob dramatically improve the robustness of EdgeGANRob. As a result, this combined method outperforms the adversarial retraining based defense method (Madry et al., 2018). (ii). Although EdgeNetRob improves the CNNs’ robustness, it decreases the clean accuracy of CNNs due to the missing texture/color information. This motivates the development of EdgeGANRob, which embeds a generative model to refill the texture/colors based on the edge images before they are fed into the classifier. Please find more visualization results on the anonymous website: <https://sites.google.com/view/edgenetrob>.

The main **contributions** of this paper include: (i) We propose a unified framework EdgeGANRob to improve the robustness of CNNs against multiple tasks simultaneously, which explicitly extracts edge/structure information from input images and then reconstructs the original images by refilling the textural information with GAN. (ii) To remain robust against sophisticated adaptive evasion attacks, in which attackers have access to the defense algorithm, we propose a robust edge detection approach *Robust Canny* based on the vanilla Canny algorithm to reduce the sensitivity of edge detector to adversarial perturbation. (iii) To further demonstrate the effectiveness of the inpainting GAN in EdgeGANRob, we also evaluate its simplified backbone procedure EdgeNetRob by performing learning tasks directly on the extracted robust edge features. To justify the above contributions, we conduct thorough evaluation on EdgeNetRob and EdgeGANRob in three tasks: adversarial attacks, distribution shifting and backdoor attacks, where significant improvements are achieved.

## 2 RELATED WORK

**Adversarial robustness** A wide range of defense methods against adversarial examples have been proposed, among which many (Liao et al., 2018; Song et al., 2018) are shown to be not robust

against adaptive attacks (Athalye et al., 2018). The state-of-the-art defense methods are based on adversarial training (Madry et al., 2018). Athalye et al. (2018) identified gradient obfuscation as a common pitfall for defense methods, thus suggested that defense methods should be evaluated against customized white-box attacks. Carlini et al. (2019) suggested that defense methods should be evaluated against strong adaptive attacks.

**Distribution shifting** Compared to adversarial examples, distribution shifting (Recht et al., 2018) is more common and general in real-world applications. Jo and Bengio (2017b) shows that CNNs have a tendency to learn the superficial statistical cues. Recently, Wang et al. (2019a) proposes a method to robustify CNNs by penalizing the predictive power of the local representations and mitigating the tendency of fitting superficial statistical cues by evaluating on four patterns, including greyscale, negcolor, radial kernel and random kernel. Hendrycks and Dietterich (2019) proposes benchmark datasets for evaluating model robustness under common perturbations.

**Backdoor attack** Backdoor attack (Chen et al., 2017a; Gu et al., 2017) is a type of poisoning attack (Shafahi et al., 2018) that works by injecting a pattern into training data. As a result, the trained models will predict a specific target class when certain pattern is injected into test data. Tran et al. (2018) has proposed a procedure to detect poisoned training data by using tools from robust statistics. Liu et al. (2018) proposes an approach to protect models from backdoor attacks by using neuron pruning.

**Robust visual features.** Recent work has highlighted a connection between recognition robustness and robust features. For image recognition, Geirhos et al. (2019); Baker et al. (2018) shows that CNNs rely more on textures than global shape structure, while humans rely more on shape structure than detailed texture. Itazuri et al. (2019) uses visualization methods and finds that adversarially robust models tend to capture global structure of the objects. Ilyas et al. (2019) argues that there exists non-robust features in natural images which are highly predictive but not interpretable by human. They showed that CNNs can obtain robustness by learning from images which contain only robust features. However, they did not directly identify which features are robust features. In this work, we propose explicitly to use edge as a robust feature.

### 3 METHODOLOGY

We introduce a new classification pipeline based on robust edge features, which we denote as EdgeGANRob. Our method first extracts edge/structure features from a given image and then reconstructs the original images by refilling the texture information with a trained generative adversarial network (GAN). The newly generated image is then fed into a classifier. In this section, we first describe a simplified backbone procedure of EdgeGANRob named EdgeNetRob, then introduce Robust Canny and inpainting GAN. Last, we describe three settings under which robustness is evaluated.

#### 3.1 EDGENETROB: A SIMPLIFIED BACKBONE OF EDGEGANROB

As a simplified backbone of EdgeGANRob, EdgeNetRob consists of two stages: First, we exploit an edge detection method to extract edge maps from an image, and then a standard image classifier  $f_\theta(\cdot)$  is trained on the extracted edge maps. Formally, denote the edge extractor function as  $e(\cdot)$ , the EdgeNetRob pipeline aims to solve the following problem:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} [\mathcal{L}(f_\theta(e(x)), y)] \quad (1)$$

where  $D$  represents the underlying data distribution and  $\mathcal{L}$  denotes the loss function (e.g., cross-entropy loss). EdgeNetRob forces the decision of CNN to be solely based on edges, thus making it less sensitive to local textures. Since original images are transformed into edge maps, even if a pre-trained classifier on the original training data is available, we still need to train the edge classifier.

Despite the simplicity of EdgeNetRob, it degrades the performance of CNNs over clean test data considering that the texture/color information is missing. This motivates us to develop EdgeGANRob which embeds a generative model to refill the texture/colors of the edge images. Because EdgeGANRob fills edge maps with texture/colors, which makes it more likely to achieve higher clean accuracy.

The robustness of such classification system that builds upon edges depends highly on the used edge detector, as many existing edge detection algorithms are also vulnerable to attacks which may lead to low accuracy of the recognition task. This motivates us to propose a robust edge detection algorithm named *Robust Canny* in the next section.

### 3.2 ROBUST EDGE DETECTION

We now describe a robust edge detection method. Note that most neural network based edge detectors are non-robust. For example, Cosgrove and Yuille (2019) finds that neural network based edge detectors like HED (Xie and Tu, 2015) can fail easily when facing adversarial perturbation. In contrast, some traditional edge detection methods like Canny (Canny, 1986) is intrinsically robust since they output binary edge maps. However, as illustrated in Figure 2 (first line), when adversarial perturbation is added, the output of Canny edge detector can become noisy. We propose to improve the robustness of vanilla Canny by truncating the noisy pixels in its intermediate stages. We refer to this modified version of Canny detector as Robust Canny.

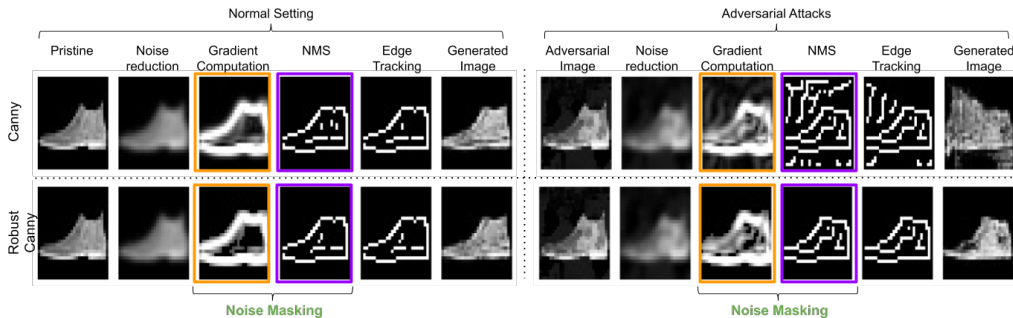


Figure 2: Visualization of intermediate stages of Vanilla Canny (*Upper*) and Robust Canny (*Lower*) on an image randomly sampled from Fashion MNIST. Results for clean images (*Left*) and adversarial images (*Right*) are presented.

The 6 stages of our proposed Robust Canny include: (1)*Noise reduction*: A Gaussian filter is applied to smooth the image. (2)*Gradient computation*: We apply the Sobel operator (Kanopoulos et al., 1988) to compute the gradient magnitude and direction at each pixel from the smoothed images. (3)*Noise masking*: We reduce the noise in the presence of adversarial perturbations by thresholding the gradient magnitudes by a level  $\alpha$ . (4)*Non-maximum suppression (NMS)*: An edge thinning step is taken to deblur the output of the Sobel operator. Gradient magnitudes that are not at a maximum along the direction of the gradient are suppressed (set to zero). (5)*Double threshold*: Using a lower threshold and a higher threshold ( $\theta_l, \theta_h$ ) for the gradient magnitude after NMS, pixels are mapped to 3 levels: strong, weak, and non-edge pixels, (6)*Edge tracking by hysteresis*: Edge pixels are detected by finding strong pixels, or weak pixels that are connected to other strong pixels. Note that we have modified the vanilla Canny algorithm by adding a noise masking stage after computing the image gradients. Later in Figure 2, we show that the gradient computation stage is sensitive to input perturbations. Thus, we set all gradient magnitudes lower than a threshold  $\alpha$  to zero to mitigate the perturbation noise. By adding a truncation operation, it is expected that adversarial noise on the gradient map with small magnitude will be reduced in early stages without affecting the quality of final edge maps. In addition to the masking operation, we find that the parameters of Canny (e.g. standard deviation of gaussian filter  $\sigma$ , thresholds  $\theta_l, \theta_h$ ) also affect the robustness level. Specifically, we notice that larger  $\sigma$  and higher thresholds  $\theta_l, \theta_h$  result in better robustness due to the stronger smoothing and pruning effects. This however comes at the cost of clean accuracy as larger  $\sigma$  leads to blurrier images and higher  $\theta_l, \theta_h$  may eliminate useful information in the output edges. To obtain a robust edge detector, we should carefully choose its parameters (e.g.,  $\sigma, \theta_l, \theta_h$ ). More details are provided in the experiment section.

### 3.3 INPAINTING GAN

In this section, we describe how we train a Generative Adversarial Network (GAN) (Goodfellow et al., 2014a) in EdgeGANRob. Recall that the task of generating color images from edge maps is

well defined under the image-to-image translation framework (pix2pix) (Isola et al., 2017). We train our inpainting GAN with two steps: in the first stage, we follow the common setup of pix2pix (Isola et al., 2017; Wang et al., 2018) to train a conditional GAN using the following objective function:

$$\min_G \max_D \mathcal{L}_{gan} = \min_G \left( \lambda_{adv} \max_D \mathcal{L}_{adv} + \lambda_{FM} \mathcal{L}_{FM} \right) \quad (2)$$

where  $\mathcal{L}_{adv}, \mathcal{L}_{FM}$  denote the adversarial loss (Goodfellow et al., 2014a) and feature matching loss (Johnson et al., 2016) with  $\lambda_{adv}$  and  $\lambda_{FM}$  control their relative importance. In the second stage, since we want our classifier to achieve high accuracy over the generated RGB images, we jointly fine-tune the trained GAN from first stage along with the classifier, using the following objective function:

$$\min_G \left( \max_D \mathcal{L}_{gan} + \lambda_{cls} \mathcal{L}_{cls} \right) \quad (3)$$

where  $\mathcal{L}_{cls}$  represents the classification loss of generated images by inpainting GAN. Note that in the first step we do not include classification loss because we want our GAN to generate more realistic images, based on which it would be easier to fine-tune the classifier to gain accuracy.

### 3.4 APPLICATIONS

Our method simultaneously improves robustness under three different settings: (i) **adversarial attack**, (ii) **distribution shifting**, (iii) **backdoor attack**. In terms of adversarial attack, EdgeGANRob is expected to improve robustness as edges are invariant to small imperceptible adversarial perturbations. Intuitively, consider a  $\ell_\infty$  threat model, it is very challenging for an attacker to make a specific edge pixel appear/disappear by reversing the magnitude of image gradient with only limited adversarial budget per pixel. When test data is under distribution shifting with well-preserved shape structure, leveraging edge features could be helpful to improve model’s generalization ability. EdgeGANRob would work in this case by focusing on shape structure which makes it less sensitive to distribution change during testing. Recall that in backdoor attack, an attacker aims to poison the training data with a specific pattern such that the trained models can be tricked into predicting a certain class when the pattern is injected at testing time. In our cases, extracting edges can be viewed as a data sanitization step to remove the malicious pattern, thus rendering potential backdoor attacks ineffective.

## 4 EXPERIMENTAL RESULTS

We evaluate the robustness of the proposed method in this section. Though EdgeNetRob is just a backbone of EdgeGANRob without inpainting GAN, our experiments show that it has unique advantage in certain settings and is of independent interest as a robust recognition method. Thus we also list it as an independent method to compare with EdgeGANRob. For our methods, we first evaluate their robustness against adversarial attacks, followed by an evaluation of their performance over distribution shifting. In addition, we evaluate the robustness against backdoor attacks.

### 4.1 EXPERIMENTAL SETUP

We conduct our experiments on two datasets: Fashion MNIST (Xiao et al., 2017) and CelebA (Liu et al., 2015). On CelebA, we evaluate our method on the task of gender classification. We did not choose the more popular MNIST and CIFAR-10 datasets because MNIST is a toy dataset where strong robustness has been achieved (Madry et al., 2018; Ding et al., 2019) and CIFAR-10 is a low-resolution dataset ( $32 \times 32$ ) where it is hard to extract semantically meaningful edges. Thus it can not provide informative benchmarks for our study. We use the same network architecture of classification for our method and the vanilla classifier. More details are shown in Appendix A.

### 4.2 ROBUSTNESS AGAINST ADVERSARIAL ATTACKS

We evaluate our methods using the commonly used  $\ell_\infty$  adversarial perturbation constraints with input range  $[0, 1]$  (Madry et al., 2018; Goodfellow et al., 2016; Song et al., 2018; Samangouei et al., 2018; Xie et al., 2019). We use standard perturbation budget on these two datasets as in Song et al. (2018); Samangouei et al. (2018); Theagarajan et al. (2019). For Fashion MNIST, we use an  $\ell_\infty$

budget of 8/256 and 25/256. For CelebA, we use an  $\ell_\infty$  budget of 2/256 and 8/256. We evaluate our methods against adaptive attack (i.e., the attacker is fully aware of the defense algorithm). Specifically, we measure the robustness to white-box attacks by the BPDA attack (Athalye et al., 2018). This attack requires a differentiable version of Canny, which is provided in Appendix C. More details on the attack setting are provided in Appendix B.

#### 4.2.1 ROBUST EDGE DETECTION

First, we illustrate why a robust edge detector is needed for defending against adversarial attacks. We compare the robustness of three edge detection methods: 1) RCF (Liu et al., 2017) which uses a CNN as backbone to generate edge maps; 2) Canny (Canny, 1986) which is the traditional Canny edge detection; 3) Robust Canny. For each of the edge detection method, we train a classifier on the extracted edge maps. The results for Fashion MNIST are reported in Table 1. First, we can see that using edges generated by RCF is not robust, as under strong adaptive attack, the accuracy drops near to 0. This result is in accordance with Cosgrove and Yuille (2019), where they show that there exists adversarial examples for neural network based edge detectors. Second, it can be noticed that adaptive attack (PGD-40) can let EdgeNetRob based on vanilla Canny drop to a low accuracy of 39.99% with perturbation  $\epsilon = 25$ . This also shows that our adaptive attacks customized for Canny is a strong adversary. Despite the weakness of vanilla Canny, we find that using Robust Canny can significantly boost the robustness under strong adaptive attack: from 39.99% to 76.75%. This shows that the truncation of values in Robust Canny is effective in reducing the adversarial risk. Therefore, for the experiments below on evaluating adversarial robustness, we use Robust Canny as the default edge extractor in EdgeNetRob and EdgeGANRob.

Table 1: Comparison of different edge extraction methods when used together with EdgeNetRob on Fashion MNIST tested with  $\epsilon = 8 / \epsilon = 25$ .

Method	Clean Accuracy	FGSM	PGD-10	PGD-40
RCF	90.15	65.79/50.07	43.68/3.37	33.84/0.18
Canny	88.32	83.45/66.98	81.24/54.07	79.76/39.99
Robust Canny	87.00	<b>84.07/79.03</b>	<b>83.88/78.53</b>	<b>83.57/76.75</b>

Table 2: Evaluation of adversarial robustness on Fashion MNIST ( $\epsilon = 8 / \epsilon = 25$ ) and CelebA ( $\epsilon = 2 / \epsilon = 8$ ).

Dataset	Method	Clean Accuracy	FGSM	PGD-10	PGD-40	CW $_\infty$
Fashion MNIST	Vanilla Net	92.88	59.50/27.82	41.55/1.76	33.35/0.48	41.82/2.00
	M-PGD	88.64/86.99	<b>85.34/78.99</b>	83.24/74.79	83.01/72.62	84.53/73.85
	EdgeNetRob	87.00	84.07/ <b>79.03</b>	83.88/ <b>78.53</b>	83.57/ <b>76.75</b>	85.53/73.43
	EdgeGANRob	87.14	85.30/78.67	<b>84.54/76.82</b>	<b>84.07/72.69</b>	<b>86.03/75.01</b>
CelebA	Vanilla Net	98.30	50.04/18.67	5.92/0.00	3.98/0.00	4.39/0.00
	M-PGD	96.51/92.75	91.73/84.67	89.01/82.55	89.01/81.31	92.46/83.45
	EdgeNetRob	94.51	93.50/87.97	93.00/84.36	92.89/82.81	93.70/83.87
	EdgeGANRob	<b>95.88</b>	<b>94.78/91.06</b>	<b>94.48/88.12</b>	<b>94.52/84.60</b>	<b>95.91/88.46</b>

#### 4.2.2 COMPARISON WITH BASELINES

We present our results for two benchmark datasets Fashion MNIST (Xiao et al., 2017) and CelebA (Liu et al., 2015). We compare with the state-of-the-art baseline: Adversarial training proposed in Madry et al. (2018). Adversarial training (Madry et al., 2018) is one of the most effective defense methods, achieving strong robustness to white-box attacks. The overall results are shown in Table 2. We notice that EdgeNetRob and EdgeGANRob leads to a small drop in clean accuracy compared to the vanilla baseline model. However, when compared with adversarial training with  $\epsilon = 8$ , both EdgeNetRob and EdgeGANRob achieve higher clean accuracy. We also observe that EdgeGANRob has higher clean accuracy than EdgeNetRob on CelebA dataset, thus validating the necessity of adding GANs on more complicated dataset to close the accuracy gap resulted from directly training on binary edge images. In terms of adversarial robustness, we observe that under

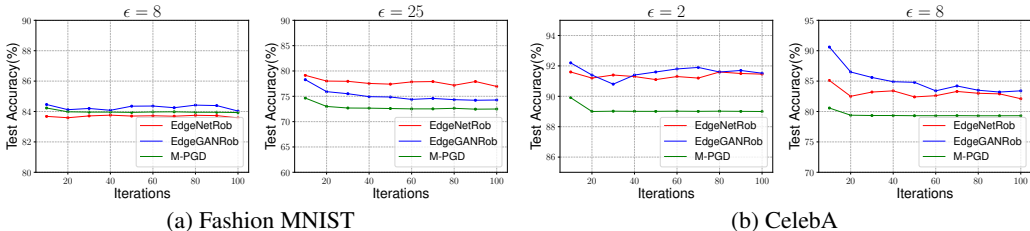


Figure 3: Test accuracy under different iterations in adaptive attacks.

strong adaptive attacks, EdgeNetRob and EdgeGANRob still remain robustness level better than or comparable to adversarial training baselines. It is worth noting that EdgeNetRob does not use adversarial training and thus has the advantage of time efficiency. We also show the plots for test accuracy under different attack iterations in Figure 3.

Table 3: Test accuracy of EdgeNetRob and EdgeGANRob on Fashion MNIST and CelebA datasets with perturbed color and texture.

Dataset	Method	Accuracy	Greyscale	NegColor	RadialKernel	RandKernel
Fashion MNIST	Vanilla Net	92.07	—	25.52	37.01	46.92
	PAR	<b>92.18</b>	—	24.20	38.67	47.94
	EdgeNetRob	88.00	—	<b>88.00</b>	<b>57.64</b>	<b>51.59</b>
	EdgeGANRob	88.57	—	<b>88.54</b>	47.77	49.87
CelebA	Vanilla Net	97.77	96.75	43.33	69.30	61.78
	PAR	<b>98.40</b>	<b>98.40</b>	59.61	73.86	61.74
	EdgeNetRob	95.02	95.02	95.02	<b>79.40</b>	74.91
	EdgeGANRob	96.28	96.28	<b>95.51</b>	77.08	<b>80.48</b>

### 4.3 ROBUSTNESS UNDER DISTRIBUTION SHIFTING

We test our method for the generalization ability under distribution shifting. We follow the experiment settings in HEX (Wang et al., 2019b) and PAR (Wang et al., 2019a), where we test the models under perturbed Fashion MNIST and CelebA with four types of patterns: greyscale, negative color, random kernel and radial kernel. The random kernel and radial kernel transformations are introduced in Jo and Bengio (2017a), which use Fourier filtering to transform an image while preserving high level semantics. We compare with state-of-the-art method PAR introduced in Wang et al. (2019a), which adds a local patchwise adversarial regularization loss. Some visualization results of perturbed images are shown in Appendix D. The overall results are shown in Table 3. We can see that EdgeNetRob and EdgeGANRob significantly improve the accuracy on three types of patterns: negative color, radial kernel and random kernel, while outperforming PAR. When testing on greyscale images, similar to baselines, our methods remain high accuracy. The results show that edge features are helpful for CNN’s generalization to test data under distribution shifting.

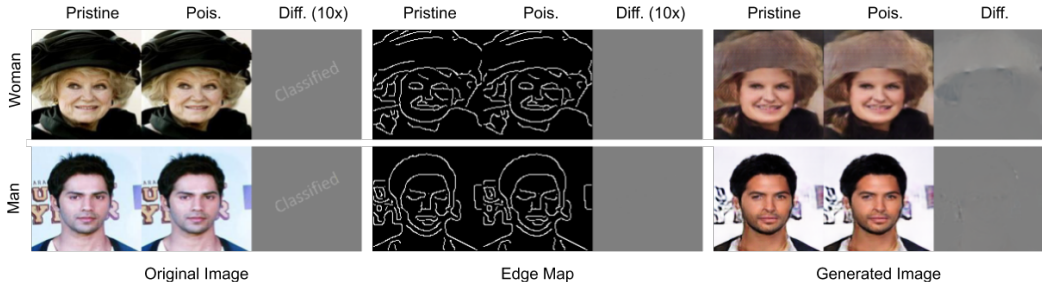


Figure 4: Qualitative results of EdgeGANRob (EdgeNetRob) with different stages for backdoor attack on CelebA.



## 4.4 ROBUSTNESS AGAINST BACKDOOR ATTACKS

We show that our method can be used as a defense against backdoor attacks. We follow the attack setup in Tran et al. (2018). We embed invisible watermark pattern letter "A" into the pristine image for Fashion MNIST and letters "classified" into CelebA. The qualitative results are shown in Figure 4 on CelebA and Figure D in Appendix for Fashion MNIST. For Fashion MNIST, we randomly choose four attack and target pairs (attack, target) as (t-shirt, trouser), (trouser, pullover), (dress, coat), (coat, dress). For CelebA, the pairs (attack, target) are (male, female) and (female, male). We select the poisoning ratio as 20% and 30% for Fashion MNIST and 5% and 10% for CelebA. We compare our method with the baseline method proposed in Tran et al. (2018), denoted as *Spectral Signature*.

The results are presented in Table 4 and Table 5, where we show the test accuracy over both standard test data ('Clean Acc') and poisoned data ('Pois Acc'). We observe that our embedding pattern can successfully attack the vanilla Net with high poisoning accuracy on both CelebA and Fashion MNIST under all settings. It can be seen that *Spectral Signature* can not always achieve good performance with such invisible watermark patterns while *EdgeNetRob* and *EdgeGANRob* consistently remain low poisoning accuracy. Figure 4 shows the qualitative results of the backdoor images after edge detection algorithm and the reconstructed images. We can observe that the effect of invisible watermark pattern can be removed by the edge detector. In addition, we find that *EdgeGANRob* achieves better clean accuracy compared with *EdgeNetRob* which also validates the benefit introduced by inserting an inpainting GAN.

Table 4: Results of *EdgeNetRob* (*EdgeGANRob*) against backdoor attack on CelebA.







Source	Target	Ratio	Method							
			Vanilla Net		<i>Spectral Signature</i>		<i>EdgeNetRob</i>		<i>EdgeGANRob</i>	
			Clean Acc	Pois Acc	Clean Acc	Pois Acc	Clean Acc	Pois Acc	Clean Acc	Pois Acc
	Man	5	98.3	97.4	<b>98.35</b>	52.89	92.3	13.80	94.53	<b>3.66</b>
		10	98.2	99.0	<b>98.03</b>	76.14	92.2	12.10	93.84	<b>5.46</b>
	Woman	5	98.2	99.0	<b>98.28</b>	93.68	94.3	<b>8.80</b>	93.91	11.73
		10	98.2	96.9	<b>98.00</b>	22.41	93.9	<b>7.70</b>	94.10	9.68

Table 5: Results of *EdgeNetRob* (*EdgeGANRob*) against backdoor attack on Fashion MNIST.

Source	Target	Ratio	Method							
			Vanilla Net		<i>Spectral Signature</i>		<i>EdgeNetRob</i>		<i>EdgeGANRob</i>	
			Clean Acc	Pois Acc	Clean Acc	Pois Acc	Clean Acc	Pois Acc	Clean Acc	Pois Acc
	T-shirt/top	20	87.17	95.80	86.32	96.30	83.10	1.00	<b>88.91</b>	<b>0.30</b>
		30	87.03	97.49	84.79	98.00	82.44	2.00	<b>88.71</b>	<b>0.30</b>
	Trouser/pants	20	86.98	93.19	87.23	93.40	82.60	<b>0.10</b>	<b>88.62</b>	1.70
		30	87.12	95.80	86.78	94.50	82.22	<b>3.90</b>	<b>88.73</b>	6.10
	Coat	20	87.55	95.59	84.88	94.80	82.90	9.2	<b>88.98</b>	<b>3.20</b>
		30	86.83	95.80	83.18	95.20	82.71	11.80	<b>88.82</b>	<b>2.92</b>
	Dress	20	86.95	90.79	87.29	10.30	82.90	9.20	<b>88.85</b>	<b>3.45</b>
		30	86.80	96.90	87.01	8.60	82.53	9.50	<b>88.38</b>	<b>3.85</b>

## 5 CONCLUSION

We introduced a new method based on robust edge features for improving general model robustness. By combining a robust edge feature extractor with the generative adversarial network, our method simultaneously achieves competitive results in terms of both adversarial robustness and generalization under distribution shifting. Additionally, we show that it can also be used to improve robustness against backdoor attacks. Our results highlight the importance of using shape information in improving model robustness and we believe it is a promising direction for future work.



## REFERENCES

- A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ICML*, 2018.
- N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.
- J. Canny. A computational approach to edge detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.
- N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017. doi: 10.1109/SP.2017.49. URL <https://doi.org/10.1109/SP.2017.49>.
- N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.
- X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017a.
- X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017b.
- C. Cosgrove and A. L. Yuille. Adversarial examples for edge detection: They exist, and they transfer. *arXiv preprint arXiv:1906.00335*, 2019.
- G. W. Ding, K. Y. C. Lui, X. Jin, L. Wang, and R. Huang. On the sensitivity of adversarial robustness to input data distributions. In *ICLR*, 2019.
- R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019.
- I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014a.
- I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.
- T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.
- A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *NIPS*, 2019.
- P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- T. Itazuri, Y. Fukuhara, H. Kataoka, and S. Morishima. What do adversarially robust models look at? *arXiv preprint arXiv:1905.07666*, 2019.
- J. Jo and Y. Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017a.
- J. Jo and Y. Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017b.

- J. Johnson, A. Alahi, and F.-F. Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- N. Kanopoulos, N. Vasanthavada, and R. L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.
- B. Landau, L. B. Smith, and S. S. Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.
- B. Li, Y. Wang, A. Singh, and Y. Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In *NIPS*, 2016.
- F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.
- K. Liu, B. Dolan-Gavitt, and S. Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. *arXiv preprint arXiv:1805.12185*, 2018.
- Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer convolutional features for edge detection. In *CVPR*, 2017.
- Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.
- P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *ICIR*, 2018.
- A. Shafahi, W. R. Huang, M. Najibi, O. Suci, C. Studer, T. Dumitras, and G. Tom. Poison frogs! targeted clean-label poisoning attacks on neural networks. *arXiv preprint arXiv:1804.00792*, 2018.
- Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018.
- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- R. Theagarajan, M. Chen, B. Bhanu, and J. Zhang. Shieldnets: Defending against adversarial attacks using probabilistic adversarial robustness. In *CVPR*, 2019.
- B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. In *NIPS*, 2018.
- H. Wang, S. Ge, Z. C. Lipton, and E. P. Xing. Learning robust global representations by penalizing local predictive power. In *NIPS*, 2019a.
- H. Wang, Z. He, Z. C. Lipton, and E. P. Xing. Learning robust representations by projecting superficial statistics out. In *ICLR*, 2019b.
- T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.
- H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- C. Xie, Y. Wu, L. v. d. Maaten, A. Yuille, and K. He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019.
- S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015.
- X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. Deep feature flow for video recognition. In *CVPR*, volume 1, page 3, 2017.

## A EXPERIMENT SETUP

For data pre-processing, we resize the images in CelebA to  $128 \times 128$  using bicubic interpolation, and use 10% of total images as test data. For both datasets, we normalize the data into the range of  $[-1, 1]$ . On Fashion-MNIST, we use a LeNet-style CNN (Table A). For CelebA dataset, we use the standard ResNet (He et al., 2016) with depth 20. Models are trained using stochastic gradient descent with momentum.

Table A: Architecture of CNN used in Fashion MNIST.

Net
Conv(128,3,3) + Relu
Conv(64,3,3) + Relu
Dropout(0.25)
FC(128) + Relu
Dropout(0.5)
FC(10) + Softmax

Table B: Hyper-parameter settings in the experiments.

Dataset	Model	Optimizer	Momentum	Epochs	Learning Rate	LR Step Decay
Fashion MNIST	LeNet	SGD	0.9	60	0.001	30, 45
CelebA	ResNet 20	SGD	0.9	40	0.1	20, 30

## B ATTACK SETTING

For each attack setting, we generate adversarial examples using three standard methods: Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014b), Projected Gradient Descent (PGD) (Madry et al., 2018) and the Carlini & Wagner  $\ell_\infty$  attack (CW) (Carlini and Wagner, 2017). For PGD attacks, we evaluate 10 steps and 40 steps PGD, denoted as ‘PGD-10’ and ‘PGD-40’ separately. For  $\ell_\infty$  distance of 2/256 or 8/256, step size is set to be 0.005. For  $\ell_\infty$  distance of 25/256, we use step size 0.015. For CW attack, we randomly sample 1,000 images for evaluation due to its high computational complexity.

We use Robust Canny for evaluation of adversarial robustness. Here we report the hyper-parameters used in Robust Canny, which are chosen using the validation set to trade off robustness and accuracy. For Fashion MNIST, we set  $\sigma = 1, \theta_l = 0.1, \theta_h = 0.2, \alpha = 0.3$ . For CelebA, we set  $\sigma = 2.5, \theta_l = 0.2, \theta_h = 0.3, \alpha = 0.2$ .

## C DIFFERENTIABLE CANNY

Note that the last three steps in the Robust Canny algorithm are non-differentiable transformations. However, in a white-box attack scenario one needs to backpropagate gradient through the edge detection algorithm for constructing adversarial samples. While obfuscating gradients through non-differentiable transformations is a commonly used defense technique, Athalye et al. (2018) show that the attacker can replace such transformation with differentiable approximations, referred to as the Backward Pass Differentiable Approximation (BPDA) technique, to construct adversarial examples. Therefore, to realize a stronger attack on our method, we find a differentiable approximation of the Robust Canny algorithm as follows.

Assuming  $x$  to hold the pixel intensities in the original image, and  $x_e$  to be the output of the Robust Canny algorithm, we can break the transformation into two stages:  $C_1(\cdot)$ , comprised of step 1-3, and  $C_2(\cdot)$  for steps 4-6 (Thresholding operation in step 3 can be formulated as a shifted ReLU function). Note that  $C_2(\cdot)$  is a non-differentiable operation, where the output is a masked version

of the input:  $C_2(x) = M(x) \otimes x$ , where  $M(\cdot)$  produces the mask (i.e., an array of zeros and ones) produced by steps 3-6, and  $\otimes$  denotes element-wise multiplication. Therefore, we can write:

$$x_e = R\text{-Canny}(x) = C_2(C_1(x)) = M(C_1(x)) \otimes C_1(x) \quad (4)$$

To obtain a differentiable approximation of R-Canny for BPDA, we assume the mask to be constant. In other words, we only backpropagate gradients through  $C_1(\cdot)$ , and not  $M(\cdot)$ .

## D MORE FIGURES IN EXPERIMENTS

In Figure A, we show the change of test accuracy under radial mask and random mask transformations with different parameters. For radial mask transformation, we vary the radius of mask in fourier domain. For random mask transformation, we sample the random masks with various probabilities. Figure B and ?? show the additional visualization results fro CelebA under four types of distribution shifting. Figure D shows the qualitative results of EdgeGANRob and EdgeNetRob for backdoor attacks on Fashion MNIST. We can also observe that the poisoning pattern can be slightly removed by EdgeNetRob and the patterns for each of the generated images do not share the similar patterns.

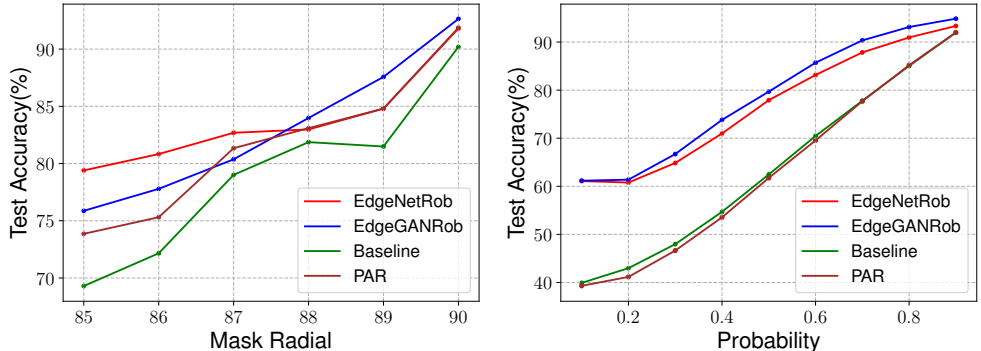


Figure A: Test accuracy under radial kernel and random kernel perturbations with different probability on CelebA.

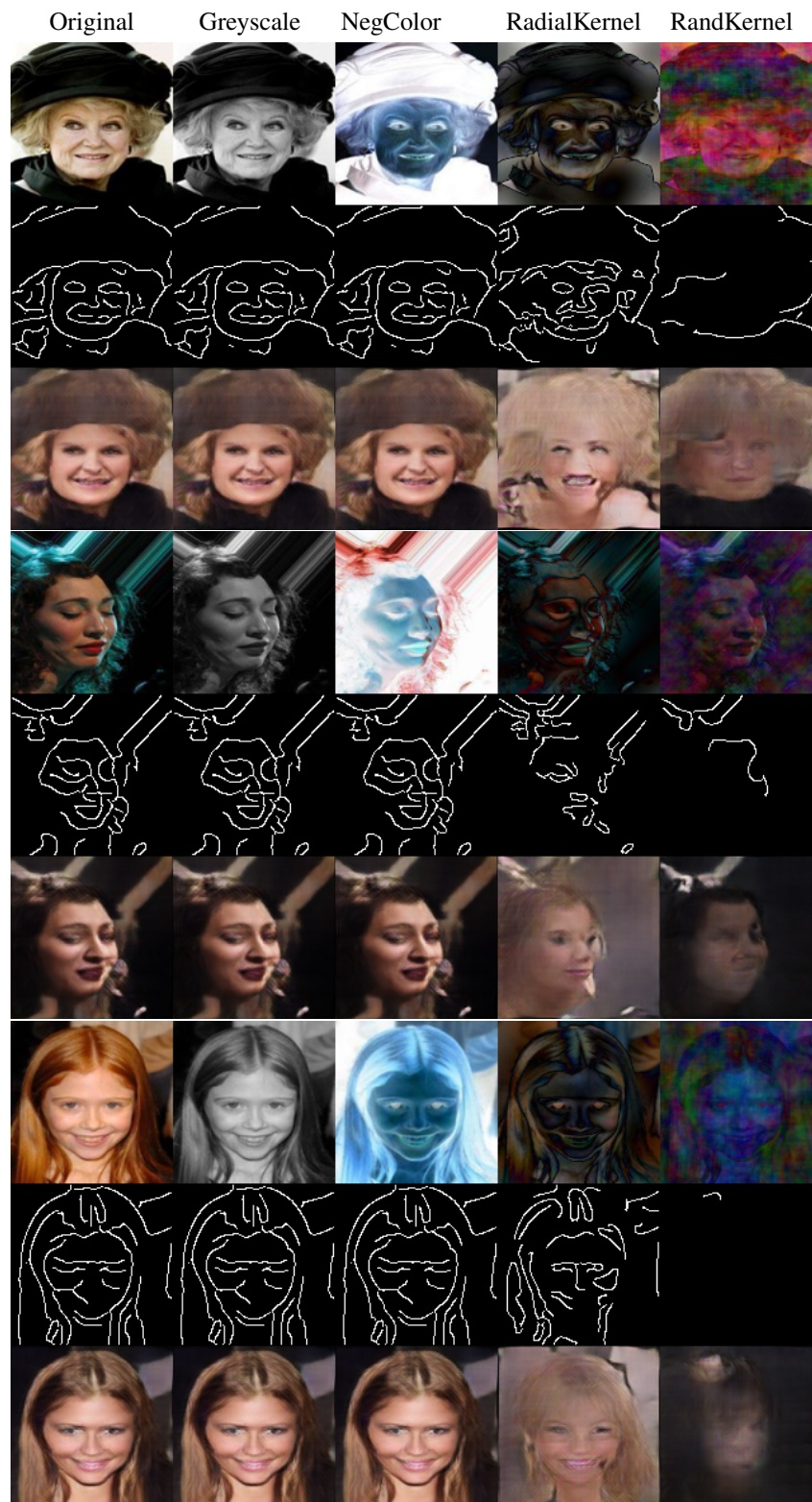


Figure B: Additional visualization of images from CelebA under four types of distribution shifting.

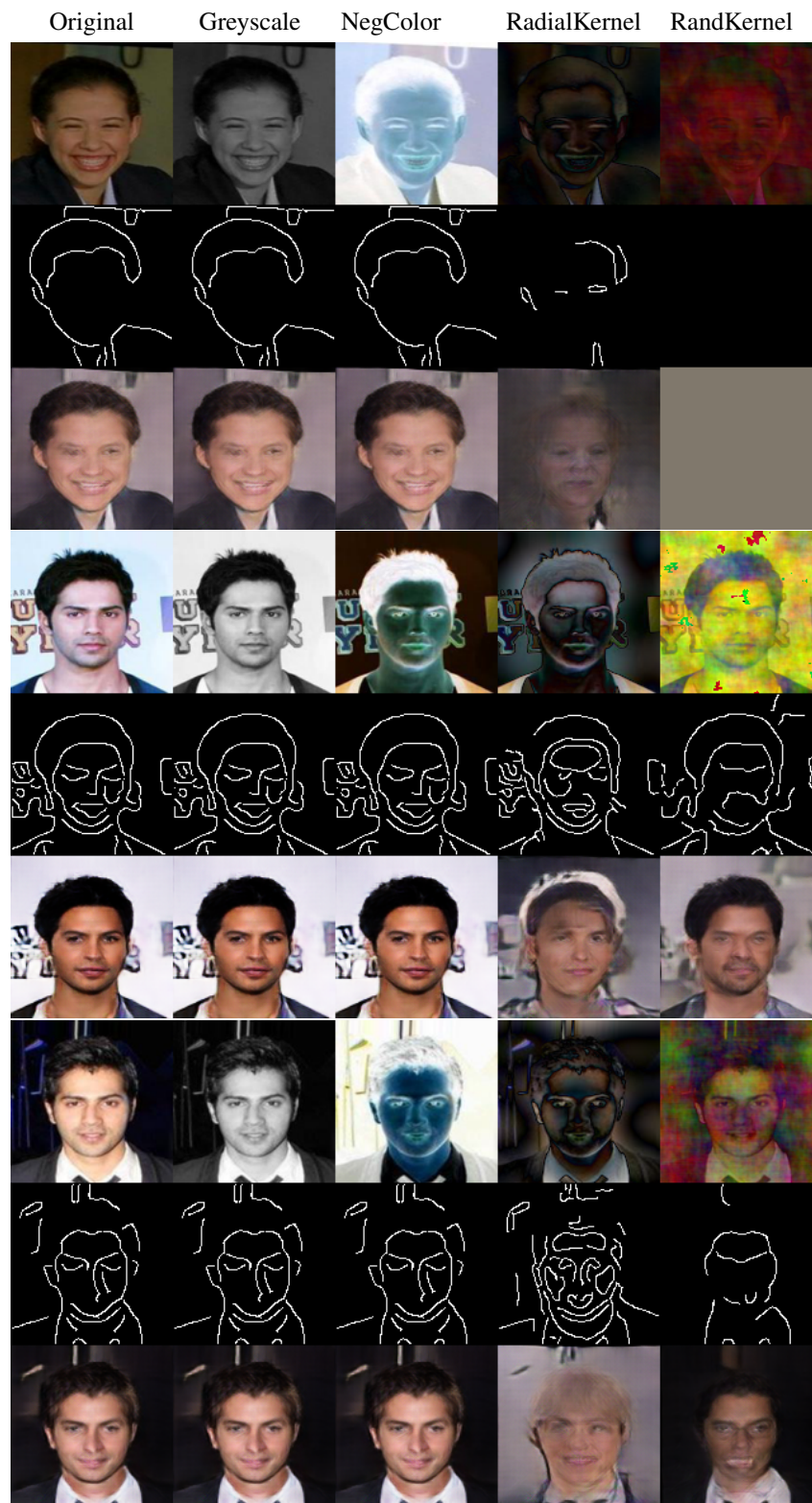


Figure C: Additional visualization of images from CelebA under four types of distribution shifting.

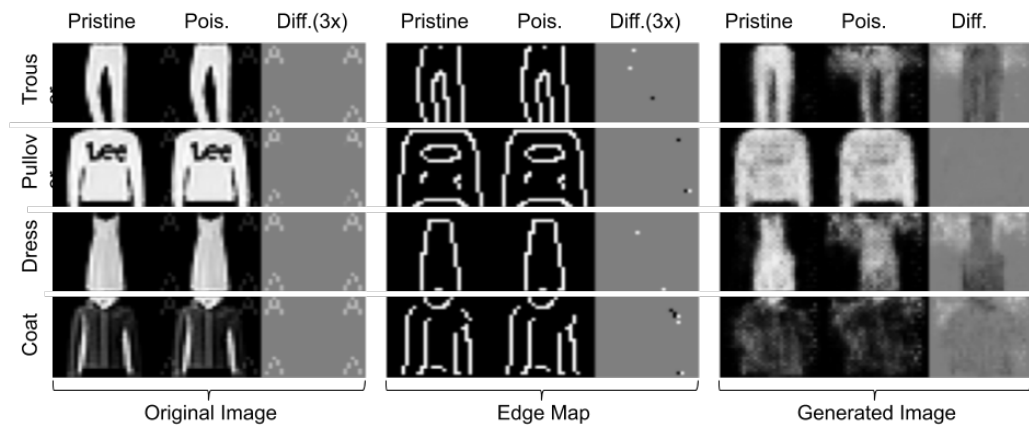


Figure D: Qualitative results of EdgeGANRob (EdgeNetRob) for backdoor attacks on Fashion MNIST.